Simulating the acquisition of object names

Alessio Plebe and **Vivian De la Cruz**
Dept. Cognitive Science
University of Messina - Italy
{alessio.plebe,vdelacruz}@unime.it

Marco Mazzone
Lab. Cognitive Science
University of Catania - Italy
mazzonem@unict.it

Abstract

Naming requires recognition. Recognition requires the ability to categorize objects and events. Infants under six months of age are capable of making fine-grained discriminations of object boundaries and three-dimensional space. At 8 to 10 months, a child's object categories are sufficiently stable and flexible to be used as the foundation for labeling and referencing actions. What mechanisms in the brain underlie the unfolding of these capacities? In this article, we describe a neural network model which attempts to simulate, in a biologically plausible way, the process by which infants learn how to recognize objects and words through exposure to visual stimuli and vocal sounds.

1 Introduction

Humans, come to recognize an infinite variety of natural and man-made objects and make use of sounds to identify and categorize them. How do human beings arrive at this capacity? Different explanations have been offered to explain the processes, and those behind the learning of first words in particular.

Evidence has made clear that object recognition and categorization in early infancy is much more sophisticated than was previously thought. By the time children are 8 to 10 months old their object categories are sufficiently stable and flexible to be used as the foundation for labeling and referencing actions. Increasing amounts of evidence point to the growing capacity of infants at this stage to reliably map arbitrary sounds onto meanings and this mapping process is crucial to the acquisition of language. The word-learning mechanisms used at this early phase of language learning could very well involve a mapping of words onto the most perceptually interesting objects in an infant's environment (Pruden et al., 2006). There are those that claim that early word learning is not purely associative and that it is based on a sensitivity to social intent (Tomasello, 1999), through joint attention phenomena (Bloom, 2000). Pruden et al. have demonstrated that 10-month-old infants "are sensitive to social cues but cannot recruit them for word learning" and therefore, at this age infants presumably have to learn words on a simple associative basis. It is not by chance, it seems, that early vocabulary is made up of the objects infants most frequently see (Gershkoff-Stowe and Smith, 2004). Early word-learning and object recognition can thus be explained, according to a growing group of researchers, by associational learning strategies alone.

There are those such as Carey and Spelke that postulate that there must necessarily be innate constraints that have the effect of making salient certain features as opposed to others, so as to narrow the hypothesis space with respect to the kinds of objects to be categorized first (Carey and Spelke, 1996). They reject the idea that object categorization in infants could emerge spontaneously from the ability to grasp patterns of statistical regularities. Jean Mandler presents evidence that the first similarity dimensions employed in categorization processes are indeed extremely general (Mandler, 2004); in other words, these dimensions single out wide domains of objects, with further refinements coming only later. Mandler claims, however, that the early salience of

Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition, pages 57–64,
Prague, Czech Republic, June 2007 ©2007 Association for Computational Linguistics

these extremely general features could have a different explanation other than nativism: for example, that salience could emerge from physiological constraints.

Using a connectionist model with backpropagation, Rogers and McClelland have shown that quite general dimensions of similarity can emerge without appealing to either physiological or cognitive constraints, simply as the result of a coherent covariation of features, that is, as an effect of mere statistical regularities (Rogers and McClelland, 2006). What Rogers and McClelland say about the most general features obviously apply also to more specific features which become salient later on. However, interesting as it is from a computational point of view, this model is rather unrealistic as a simulation of biological categorization processes.

Linda Smith, suggests that words can contribute to category formation, in that they behave as features which co-vary with other language-independent features of objects (Smith, 1999). In general, her idea is that the relevant features simply emerge from regularities in the input. Terry Regier, building upon the proposal offered by Smith, has shown that word learning might behave in analogy with what we have said about categorization (Regier, 2005): certain features of both objects and words (i.e., phonological forms) can be made more salient than others, simply as a consequence of regularities in objects, words, and their co-variation. Regier's training sets however, are constituted by wholly "artificial phonological or semantic features", rather than by "natural features such as voicing or shape". The positions mentioned above conflict with others, such as that of Lila Gleitman and her colleagues, according to which some innate constraints are needed in order to learn words. It should be noted, however, that even in Gleitman's proposal the need for innate constraints on syntax-semantic mapping mainly concerns verbs; moreover, the possibility to apprehend a core set of concrete terms without the contribution of any syntactic constraint is considered as a precondition for verb acquisition itself (Gillette et al., 1999).

This paper describes a neural network model which attempts to simulate the process by which infants learn how to recognize objects and words in the first year of life through exposure to visual stimuli and vocal sounds. The approach here pursued is in line with the view that a coherent covariation of features is the major engine leading to object name acquisition, the attempt made however, is to rely on biological ways of capturing coherent covariation. The pre-established design of the mature functions of the organism is avoided, and the emergence of the final function of each component of the system is left to the plastic development of the neural circuits. In the cortex, there is very little differentiation in the computational capability that neural circuits will potentially perform in the mature stage. The interaction between environmental stimuli and some of the basic mechanisms of development is what drives differentiation in computational functions. This position has large empirical support (Katz and Callaway, 1992; Löwel and Singer, 2002), and is compatible with current knowledge on neural genetics (Quartz, 2003).

The model here described, can be considered an implementation of the processes that emerge around the 10 month of age period. It can also be used to consider what happens in a hypothesized subsequent period, in which the phenomenon of joint attention provides the social cueing that leads to the increased ability to focus on certain objects as opposed to others.

2 The proposed model

First the mathematics common to the modules will be described, then the model will be outlined. Details of the visual and the auditory paths will be provided along with a description of the learning procedures.

2.1 The mathematical abstraction of the cortical maps

All the modules composing this model are implemented as artificial cortical maps, adopting the LISSOM (*Laterally Interconnected Synergetically Self-Organizing Map*) architecture (Sirosh and Miikkulainen, 1997; Bednar, 2002). This architecture has been chosen because of its reproduction of neural plasticity, through the combination of Hebb's principle and neural homeostasis, and because it is a good compromise between a number of realistic features and the simplicity necessary for building complex

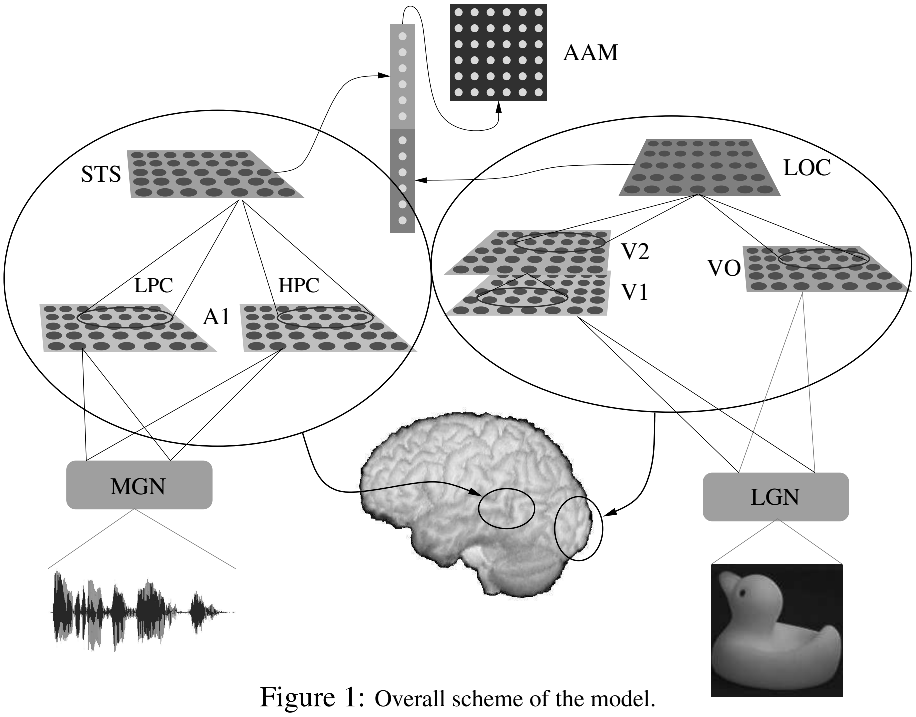

Figure 1: Overall scheme of the model.

models. The LISSOM is a two dimensional arrangement of neurons, where each cell is not only connected with the afferent input vector, but receives excitatory and inhibitory inputs from several neighbor neurons on the same map:

$$
\begin{aligned}
x_i^{(k)} = f \Bigg(& \frac{\gamma_{\mathrm{A}}}{1+\gamma_{\mathrm{N}} \vec{I} \cdot \vec{v}_{r_{\mathrm{A}},i}} \vec{a}_{r_{\mathrm{A}},i} \cdot \vec{v}_{r_{\mathrm{A}},i} \\
& + \gamma_{\mathrm{E}} \vec{e}_{r_{\mathrm{E}},i} \cdot \vec{x}_{r_{\mathrm{E}},i}^{(k-1)} \\
& - \gamma_{\mathrm{H}} \vec{h}_{r_{\mathrm{H}},i} \cdot \vec{x}_{r_{\mathrm{H}},i}^{(k-1)} \Bigg),
\end{aligned} \tag{1}
$$

where $x_i^{(k)}$ is the activation of the neuron i at time step k. All vectors are composed by a circular neighborhood of given radius around the neuron i: vectors $\vec{x}^{\,(k-1)}$ are activations of neurons on the same layer at the previous time step. Vector $\vec{v}_{r_{\mathrm{A}},i}$ comprises all neurons in the underlying layer, in a circular area centered on the projection of i on this layer, with radius r_{A}. Vectors $\vec{a}_{r_{\mathrm{A}},i}$, $\vec{e}_{r_{\mathrm{E}},i}$, and $\vec{h}_{r_{\mathrm{H}},i}$ are composed by all connection strengths of, afferent, excitatory or inhibitory neurons respectively, projecting to i, inside circular areas of radius r_{A}, r_{E}, r_{H}. Vector $\vec{I}$ is just a vector of 1's of the same dimension of $\vec{v}_{r_{\mathrm{A}},i}$. The scalars γ_{A}, γ_{E}, and γ_{H}, are constants modulating the contribution of afferent, excitatory and inhibitory connections. The scalar γ_{N} controls the setting of a push-pull effect in the afferent weights, allowing inhibitory effect without negative weight values. Mathematically, it represents dividing the response from the excitatory weights by the response from a uniform disc of inhibitory weights over the receptive field of neuron i. The map is characterized by the matrices $\mathbf{A}, \mathbf{E}, \mathbf{H}$, which columns are all vectors $\vec{a}$, $\vec{e}$, $\vec{h}$ for every neuron in the map. The function f is a monotonic non-linear function limited between 0 and 1. The final activation value of the neurons is assessed after settling time K.

All connection strengths to neuron i adapt by following the rules:

$$
\Delta \vec{a}_{r_{\mathrm{A}},i} = \frac{\vec{a}_{r_{\mathrm{A}},i} + \eta_{\mathrm{A}} x_i \vec{v}_{r_{\mathrm{A}},i}}{\left\| \vec{a}_{r_{\mathrm{A}},i} + \eta_{\mathrm{A}} x_i \vec{v}_{r_{\mathrm{A}},i} \right\|} - \vec{a}_{r_{\mathrm{A}},i}, \tag{2}
$$

$$
\Delta \vec{e}_{r_{\mathrm{E}},i} = \frac{\vec{e}_{r_{\mathrm{E}},i} + \eta_{\mathrm{E}} x_i \vec{x}_{r_{\mathrm{E}},i}}{\left\| \vec{a}_{r_{\mathrm{E}},i} + \eta_{\mathrm{E}} x_i \vec{x}_{r_{\mathrm{E}},i} \right\|} - \vec{e}_{r_{\mathrm{E}},i}, \tag{3}
$$

$$
\Delta \vec{h}_{r_{\mathrm{H}},i} = \frac{\vec{h}_{r_{\mathrm{H}},i} + \eta_{\mathrm{A}} x_i \vec{x}_{r_{\mathrm{H}},i}}{\left\| \vec{h}_{r_{\mathrm{H}},i} + \eta_{\mathrm{A}} x_i \vec{x}_{r_{\mathrm{H}},i} \right\|} - \vec{h}_{r_{\mathrm{H}},i}, \tag{4}
$$

where $\eta_{\{\mathrm{A},\mathrm{E},\mathrm{H}\}}$ are the learning rates for afferent, excitatory and inhibitory synaptic modifications. All rules are based on the Hebbian law, with an additional competitive factor, here implemented as a normalization, that maintains constant the integration of all connection strengths to the same neu-

	layer	size	r_{A}	r_{E}	r_{H}	γ_{A}	γ_{E}	γ_{H}	γ_{N}
LGN	Lateral Geniculated Nucleus	120×120	-	-	-	-	-	-	-
MGN	Medial Geniculated Nucleus	32×32	-	-	-	-	-	-	-
V1	Primary Visual Cortex	96×96	8.5	1.5	7.0	1.5	1.0	1.0	0.0
V2	Secondary Visual Cortex	30×30	7.5	8.5	3.5	50.0	3.2	2.5	0.7
VO	Ventral Occipital	30×30	24.5	4.0	8.0	1.8	1.0	1.0	0.0
A1	Auditory Primary Cortex	24×24	3.5	2.5	5.5	5.0	5.0	6.7	0.8
LOC	Lateral Occipital Complex	16×16	6.5	1.5	3.5	1.2	1.0	1.5	0.0
STS	Superior Temporal Sulcus	16×16	3.5	2.5	2.5	2.0	1.6	2.6	0.0

Table 1: Legend of all modules, and main parameters of the cortical layers composing the model.

ron, and to the same type (afferent, excitatory or inhibitory). This is a computational account of the biological phenomena of homeostatic plasticity, that induce neurons in the cortex to maintain an average firing rate by correcting their incoming synaptic strengths.

2.2 The overall model

An outline of the modules that make up the model is shown in Fig. 1. The component names and their dimensions are in Tab. 1. All cortical layers are implemented by LISSOM maps, where the afferent connections $\vec{v}$ in (1) are either neurons of lower LISSOM maps, or neurons in the thalamic nuclei MGN and LGN. There are two main paths, one for the visual process and another for the auditory channel. Both paths include thalamic modules, which are not the object of this study and are therefore hardwired according to what is known about their functions. The two higher cortical maps, LOC and STS, will carry the best representation coded by models on object visual features and word features. These two representations are associated in an abstract type map, called AAM (*Abstract Associative Map*). This component is implemented using the SOM (*Self Organized Map*) (Kohonen, 1995) architecture, known to provide non linear bidimensional ordering of input vectors by unsupervised mechanisms. It is the only component of the model that cannot be conceptually referred to as a precise cortical area. It is an abstraction of processes that actually involve several brain areas in a complex way, and as such departs computationally from realistic cortical architecture.

2.3 The visual pathway

As shown in Fig. 1, the architecture here used includes hardwired extracortical maps with simple on-center and off-center receptive fields. There are three pairs of sheets in the LGN maps: one connected to the intensity image plane, and the other two connected to the medium and long wavelength planes. In the color channels the internal excitatory portion of the receptive field is connected to the channel of one color, and the surrounding inhibitory part to the opposite color. The cortical process proceeds along two different streams: the achromatic component is connected to the primary visual map V1 followed by V2, the two spectral components are processed by map VO, the color center, also called hV4 or V8 (Brewer et al., 2005). The two streams rejoin in the cortical map LOC, the area recently suggested as being the first involved in object recognition in humans (Malach et al., 1995; Kanwisher, 2003). Details of the visual path are in (Plebe and Domenella, 2006).

2.4 The auditory pathway

The hardwired extracortical MGN component is just a placeholder for the spectrogram representation of the sound pressure waves, which is extracted with tools of the *Festival* software (Black and Taylor, 1997). It is justified by evidence of the spectro-temporal process performed by the cochlear-thalamic circuits (Escabi and Read, 2003). The auditory primary cortex is simulated by a double sheet of neurons, taking into account a double population of cells found in this area (Atzori et al., 2001), where the so-called LPC (*Low-Probability Connections*) is sensitive to the stationary component of the sound signal and the HPC (*High-Probability Connections*) population responds to transient inputs mainly. The next map in the auditory path of the model is STS, because the superior temporal sulcus is believed to be the main brain area responsive to vocal sounds (Belin et al., 2002).

2.5 The Abstract Associative Map

The upper AAM map in the model reflects how the system associates certain sound forms with the visual appearance of objects, and has the main purpose of showing what has been achieved in the cortical part of the model. It is trained using the outputs of the STS and the LOC maps of the model. After training, each neuron x in AAM is labeled, according to different test conditions X. The labeling function $l(\cdot)$ associates the neuron x with an entity e, which can be an object o of the COIL set $\mathcal{O}$, when $X \in \{\mathrm{A}, \mathrm{B}\}$ or a category c of the set $\mathcal{C}$ for the test condition $X \in \{\mathrm{C}, \mathrm{D}\}$. The general form of the labeling function is:

$$l^{(X)}(x) = \arg\max_{e \in \mathcal{E}} \left\{ \left| \mathcal{W}_x^{(e)} \right| \right\} \qquad (5)$$

where $\mathcal{W}_x^{(e)}$ is a set of sensorial stimuli related to the element $e \in \mathcal{E}$, such that their processing in the model activate x as winner in the AMM map. The set $\mathcal{E}$ can be $\mathcal{O}$ or $\mathcal{C}$ depending on X. The neuron x elicited in the AAM map as the consequence of presenting a visual stimulus v_o of an object o and a sound stimulus s_c of a tagory c is given by the function $x = w(v_o, s_c)$ with the convention that $w(v, \epsilon)$ computes the winning neuron in AAM comparing only the LOC portion of the coding vector, and $w(\epsilon, s)$ only the STS portion. The function $b(o) : \mathcal{O} \rightarrow \mathcal{C}$ associates an object o to its category. Here four testing conditions are used:

- A object recognition by vision and audio
- B object recognition by vision only
- C category recognition by vision and audio
- D category recognition by audio only

corresponding to the following $\mathcal{W}$ sets in (5):

$$\mathrm{A} : \left\{ v_o : x = w(v_o, s_{c(o)}) \right\} \qquad (6)$$

$$\mathrm{B} : \{ v_o : x = w(v_o, \epsilon) \} \qquad (7)$$

$$\mathrm{C} : \{ v_o : c = b(o) \wedge x = w(\epsilon, s_c) \} \qquad (8)$$

$$\mathrm{D} : \{ s_c : x = w(\epsilon, s_c) \} \qquad (9)$$

From the labeling functions the possibility of estimating the accuracy of recognition immediately follows, simply by weighing the number of cases where the category or the object has been classified as the prevailing one in each neuron of the AAM SOM.

2.6 Exposure to stimuli

The visual path in the model develops in two stages. Initially the inputs to the network are synthetic random blobs, simulating pre-natal waves of spontaneous activity, known to be essential in the early development of the visual system (Sengpiel and Kind, 2002). In the second stage, corresponding to the period after eye opening, natural images are used. In order to address one of the main problems in recognition, the identifying of an object under different views, the COIL-100 collection has been used (Nayar and Murase, 1995) where 72 different views are available for each of the 100 objects. Using natural images where there is only one main object is cleary a simplification in the vision process of this model, but it does not compromise the realism of the conditions. It always could be assumed that the single object analysis corresponds to a foval focusing as consequence of a saccadic move, cued by any attentive mechanism.

In the auditory path there are different stages as well. Initially, the maps are exposed to random patches in frequency-time domain, with shorter duration for HPC and longer for LPC. Subsequently, all the auditory maps are exposed to the 7200 most common English words (from `http://www.bckelk.uklinux.net/menu.html`) with lengths between 3 and 10 characters. All words are converted from text to waves using *Festival* (Black and Taylor, 1997), with cepstral order 64 and a unified time window of 2.3 seconds. Eventually, the last stage of training simulates events when an object is viewed and a word corresponding to its basic category is heard simultaneously. The 100 objects have been grouped manually into 38 categories. Some categories, such as `cup` or `medicine` count 5 exemplars in the object collection, while others, such as `telephone`, have only one exemplar.

3 Results

3.1 Developed functions in the cortical maps

At the end of development each map in the model has evolved its own function. Different functions have emerged from identical computational architectures. The differences are due to the different positions of a maps in the modules hierarchy, to different exposure to environmental stimuli, and different structural parameters. The functions obtained in the experiment are the following. In the visual path orientation selectivity emerged in the model's V1 map as demonstrated in (Sirosh and Miikkulainen, 1997) and (Plebe and Domenella, 2006). Orientation selectivity is the main organization in primary visual cortex, where the responsiveness of neurons to oriented segments is arranged over repeated patterns of gradually changing orientations, broken by few discontinuities (Vanduffel et al., 2002). Angle selectivity emerged in the model's V2 map. In the secondary visual cortex the main recently discovered phenomena is the selectivity to angles (Ito and Komatsu, 2004), especially in the range between 60 and 150 degrees. The essential features of color constancy are reproduced in the model's VO map, which is the ability of neurons to respond to specific hues, regardless of intensity. Color constancy is the tendency of the color of a surface to appear more constant that it is in reality. This property is helpful in object recognition, and develops sometime between two and four months of age. (Dannemiller, 1989). One of the main functions shown by the LOC layer in the model is visual invariance, the property of neurons to respond to peculiar object features despite changes in the object's appearance due to different points of view. Invariance indeed is one of the main requirements for an object-recognition area, and is found in human LOC (Grill-Spector et al., 2001; Kanwisher, 2003). Tonotopic mapping is a known feature of the primary auditory cortex that represents the dimensions of frequency and time sequences in a sound pattern (Verkindt et al., 1995). In the model it is split into a sheet where neurons have receptive fields that are more elongated along the time dimension (LPC) and another where the resulting receptive fields are more elongated along the frequency dimension (HPC). The spectrotemporal mapping obtained in STS is a population coding of features, in frequency and time domains, representative of the sound patterns heard during the development phase. It therefore reflects the statistical phonemic regularities in common spoken English, extracted from the 7200 training samples.

category	test A	test B	test C	test D
medicine	0.906	0.803	1.0	1.0
fruit	1.0	0.759	1.0	1.0
boat	0.604	0.401	1.0	1.0
tomato	1.0	0.889	1.0	1.0
sauce	1.0	1.0	1.0	1.0
car	0.607	0.512	0.992	1.0
drink	0.826	0.812	1.0	1.0
soap	0.696	0.667	1.0	1.0
cup	1.0	0.919	1.0	0.0
piece	0.633	0.561	1.0	1.0
kitten	1.0	0.806	1.0	1.0
bird	1.0	1.0	1.0	1.0
truck	0.879	0.556	1.0	1.0
dummy	1.0	0.833	1.0	1.0
tool	0.722	0.375	1.0	1.0
pottery	1.0	1.0	1.0	1.0
jam	1.0	1.0	1.0	1.0
frog	1.0	0.806	1.0	1.0
cheese	0.958	0.949	1.0	1.0
bottle	0.856	0.839	1.0	1.0
hanger	1.0	0.694	1.0	1.0
sweets	1.0	0.701	1.0	1.0
tape	1.0	0.861	1.0	1.0
mug	0.944	0.889	1.0	1.0
spoon	1.0	0.680	1.0	1.0
cigarettes	0.972	0.729	0.972	1.0
ring	1.0	1.0	1.0	1.0
pig	1.0	0.778	1.0	1.0
dog	1.0	0.917	1.0	1.0
toast	1.0	0.868	1.0	1.0
plug	1.0	0.771	1.0	1.0
pot	1.0	0.681	1.0	1.0
telephone	1.0	0.306	1.0	1.0
pepper	1.0	0.951	1.0	1.0
chewinggum	0.954	0.509	1.0	1.0
chicken	1.0	0.944	1.0	1.0
jug	1.0	0.917	1.0	1.0
can	1.0	0.903	1.0	1.0

Table 2: Accuracy in recognition measured by labeling in the AAM, for objects grouped by category.

3.2 Recognition and categorization in AAM

The accuracy of object and category recognition under several conditions is shown in Table 2. All tests clearly prove that the system has learned an efficient capacity of object recognition and naming, with respect to the small world of object and names used in the experiment. Tests C and D demonstrate that the recognition of categories by names is almost complete, both when hearing a name or when seeing an object and hearing its name. In tests A and B, the recognition of individual objects is also very high. In several cases, it can be seen that names also help in the recognition of individual objects. One of the clearest cases is the category `tool` (shown in Fig. 2),

shape	test A	test B	Δ
h-parallelepiped	0.921	0.712	0.209
round	1.0	0.904	0.096
composed	0.702	0.565	0.137
q-cylindrical	0.884	0.861	0.023
q-h-parallelepiped	0.734	0.513	0.221
cylindrical	0.926	0.907	0.019
cup-shaped	0.975	0.897	0.078
q-v-parallelepiped	0.869	0.754	0.115
body	1.0	0.869	0.131
conic	1.0	1.0	0.0
parallelepiped	0.722	0.510	0.212
q-parallelepiped	1.0	0.634	0.366

Table 3: Accuracy in recognition measured by labeling in the AAM, for objects grouped by their visual shape, Δ is the improvement gained with naming.

where the accuracy for each individual object doubles when using names. It seems to be analogous to the situation described in (Smith, 1999), where the word contributes to the emergence of patterns of regularity. The 100% accuracy for the category, in this case, is better accounted for as an emergent example of synonymy, where coupling with the same word is accepted, despite the difference in the output of the visual process.

In table 3 accuracy results for individual objects are listed, grouped by object shape. In this case category accuracy cannot be computed, because shapes cross category boundaries. It can be seen that the improvement Δ is proportional to the salience in shape: it is meaningless for common, obvious shapes, and higher when object shape is uncommon. This result is in agreement with findings in (Gershkoff-Stowe and Smith, 2004).

4 Conclusions

The model here described attempts to simulate lexical acquisition from auditory and visual stimuli from a brain processes point of view. It models these processes in a biologically plausible way in that it does not begin with a predetermined design of mature functions, but instead allows final functions of the components to emerge as a result of the plastic development of neural circuits. It grounds this choice and its design principles in what is known of the cerebral cortex. In this model, the overall important result achieved so far, is the emergence of naming and recognition abilities exclusively through exposure of the system to environmental stimuli, in terms of activities similar to pre-natal spontaneous activities, and later to natural images and vocal sounds. This result has interesting theoretical implications for developmental psychologists and may provide a useful computational tool for future investigations on phenomena such as the effects of shape on object recognition and naming.

In conclusion this model represents a first step in simulating the interaction of the visual and the auditory cortex in learning object recognition and naming, and being a model of high level complex cognitive functions, it necessarily lacks several details of the biological cortical circuits. It lacks biological plausibility in the auditory path because of the state of current knowledge of the processes going on there. Future developments of the model will foresee the inclusion of backprojections between maps in the hierarchy, trials on preliminary categorization at the level of phonemes and syllables in the auditory path, as well as exposure to images with multiple objects in the scene.

References

Marco Atzori, Saobo Lei, D. Ieuan P. Evans, Patrick O. Kanold, Emily Phillips-Tansey, Orinthal McIntyre, and Chris J. McBain. 2001. Differential synaptic processing separates stationary from transient inputs to the auditory cortex. *Neural Networks*, 4:1230–1237.

James A. Bednar. 2002. *Learning to See: Genetic and Environmental Influences on Visual Development*. Ph.D. thesis, University of Texas at Austin. Tech Report AI-TR-02-294.

Pascal Belin, Robert J. Zatorre, and Pierre Ahad. 2002. Human temporal-lobe response to vocal sounds. *Cognitive Brain Research*, 13:17–26.

Alan W. Black and Paul A. Taylor. 1997. The festival speech synthesis system: System documentation. Technical Report HCRC/TR-83, Human Communciation Research Centre, University of Edinburgh, Edinburgh, UK.

Paul Bloom. 2000. *How children learn the meanings of words*. MIT Press, Cambridge (MA).

Alyssa A. Brewer, Junjie Liu, Alex R. Wade, and Brian A. Wandell. 2005. Visual field maps and stimulus selectivity in human ventral occipital cortex. *Nature Neuroscience*, 8:1102–1109.

Susan Carey and Elizabeth Spelke. 1996. Science and core knowledge. *Journal of Philosophy of Science*, 63:515–533.

James L. Dannemiller. 1989. A test of color constancy in 9- and 20-weeks-old human infants following simulated illuminant changes. *Developmental Psychology*, 25:171–184.

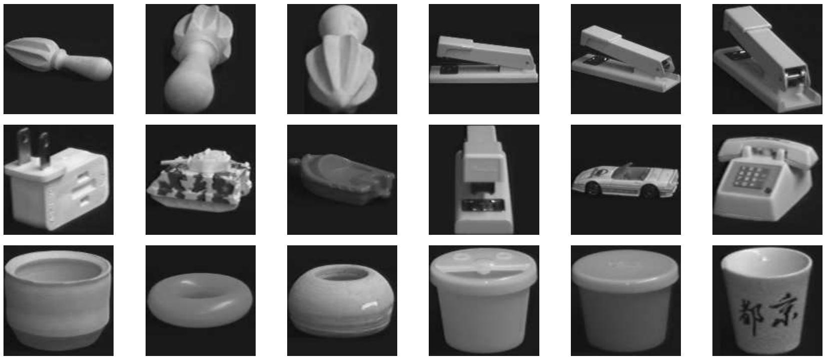

Figure 2: Objects mentioned in the discussion on recognition results. In the upper row views of the two objects of the category `tool`. In the middle row objects with difficult shapes (`q-h-parallelepiped`, `q-parallelepiped`). In the lower row objects with easy shapes (`cylindrical`, `round`, and `conic`).

Monty A. Escabi and Heather L. Read. 2003. Representation of spectrotemporal sound information in the ascending auditory pathway. *Biological Cybernetics*, 89:350–362.

Lisa Gershkoff-Stowe and Linda B. Smith. 2004. Shape and the first hundred nouns. *Child Development*, 75:1098–1114.

Jane Gillette, Henry Gleitman, Lila Gleitman, and Anne Lederer. 1999. Human simulations of vocabulary learning. *Cognition*, 73:135–176.

Kalanit Grill-Spector, Zoe Kourtzi, and Nancy Kanwisher. 2001. The lateral occipital complex and its role in object recognition. *Vision Research*, 41:1409–1422.

Minami Ito and Hidehiko Komatsu. 2004. Representation of angles embedded within contour stimuli in area V2 of macaque monkeys. *Journal of Neuroscience*, 24:3313–3324.

Nancy Kanwisher. 2003. The ventral visual object pathway in humans: Evidence from fMRI. In Leo Chalupa and John Werner, editors, *The Visual Neurosciences*. MIT Press, Cambridge (MA).

Lawrence C. Katz and Edward M. Callaway. 1992. Development of local circuits in mammalian visual cortex. *Annual Review Neuroscience*, 15:31–56.

Teuvo Kohonen. 1995. *Self-Organizing Maps*. Springer-Verlag, Berlin.

Siegrid Löwel and Wolf Singer. 2002. Experience-dependent plasticity of intracortical connections. In Manfred Fahle and Tomaso Poggio, editors, *Perceptual Learning*. MIT Press, Cambridge (MA).

R. Malach, J. B. Reppas, R. R. Benson, K. K. Kwong, H. Jiang, W. A. Kennedy, P. J. Ledden, T. J. Brady, B. R. Rosen, and R. B.H. Tootell. 1995. Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proceedings of the Natural Academy of Science USA*, 92:8135–8139.

Jean Matter Mandler. 2004. *The Foundations of Mind*. Oxford University Press, Oxford (UK).

Shree Nayar and Hiroshi Murase. 1995. Visual learning and recognition of 3-d object by appearence. *International Journal of Computer Vision*, 14:5–24.

Alessio Plebe and Rosaria Grazia Domenella. 2006. Early development of visual recognition. *BioSystems*, 86:63–74.

Shannon M. Pruden, Kathy Hirsh-Pasek, Roberta Michnick Golinkoff, and Elizabeth A. Hennon. 2006. The birth of words: Ten-month-olds learn words through perceptual salience. *Child Development*, 77:266–280.

Steven R. Quartz. 2003. Innateness and the brain. *Biology and Philosophy*, 18:13–40.

Terry Regier. 2005. The emergence of words: Attentional learning in form and meaning. *Cognitive Science*, 29:819–865.

Timothy T. Rogers and James L. McClelland. 2006. *Semantic Cognition - A Parallel Distributed Processing Approach*. MIT Press, Cambridge (MA).

Frank Sengpiel and Peter C. Kind. 2002. The role of activity in development of the visual system. *Current Biology*, 12:818–826.

Joseph Sirosh and Risto Miikkulainen. 1997. Topographic receptive fields and patterned lateral interaction in a self-organizing model of the primary visual cortex. *Neural Computation*, 9:577–594.

Linda B. Smith. 1999. Children's noun learning: How general learning processes make specialized learning mechanisms. In Brian MacWhinney, editor, *The Emergence of Language*. Lawrence Erlbaum Associates, Mahwah (NJ). Second Edition.

Michael Tomasello. 1999. *The cultural origins of human cognition*. Harvard University Press, Cambridge (MA).

Wim Vanduffel, Roger B.H. Tootell, Anick A. Schoups, and Guy A. Orban. 2002. The organization of orientation selectivity throughout the macaque visual cortex. *Cerebral Cortex*, 12:647–662.

Chantal Verkindt, Olivier Bertrand, Franǫis Echallier, and Jacques Pernier. 1995. Tonotopic organization of the human auditory cortex: N100 topography and multiple dipole model analysis. *Electroencephalography and Clinical Neurophisiology*, 96:143–156.